Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240697

Comprehensive analysis of pathological and biochemical data of prostate tumors based on multiple machine learning models

Ziwen Wan[1], Lehan Zou[1], Junjie Shen[1], Wenxi Teng[1], Shuang Wang[2], Zhibin Zhao[1,3,*]

[1]Nanchang University, 235 Nanjing East Road, Nanchang 330000, China
[2]Jiangxi University of Finance and Economics, 169 Shuanggang East Street, Nanchang, Jiangxi, 330000, China

[3]zhaozhibin@ncu.edu.cn
*corresponding author

Abstract. This study conducts a comprehensive analysis of global prostate tumor cases, employing various machine learning models for classification. The research focuses on data preprocessing, feature selection, and model training, utilizing effective techniques for handling biochemical and pathological data. In model training, a range of algorithms such as XGBoost, Random Forest, LightGBM, and CatBoost were explored, ultimately achieving an 89% test accuracy through a Stacking model. For pathological data, strategies included addressing missing values, text cleaning, and segmentation, combining CountVectorizer with multiple algorithms. LightGBM demonstrated optimal performance with an accuracy of 82%. The Bert model, post-segmentation and Focal loss optimization, achieved an accuracy of 80.39%. In summary, this research leverages machine learning to provide a convenient approach for diagnosing and treating prostate tumors, offering robust model support for predicting related diseases.

Keywords: Prostate neoplasms, Machine learning, Biochemical and pathological data, Feature selection.

1. Introduction

Prostate cancer is one of the major public health challenges for men worldwide. Statistics show that it is one of the most common tumors in men and the leading cause of cancer-related death [1]. At the same time, benign prostatic hyperplasia (BPH) is also prevalent in the elderly male population [2].BPH is considered an important risk factor for certain urologic cancers and is particularly relevant for prostate cancer. Global cancer data in 2020 showed that there were 1.41 million new cases of prostate cancer, making it the second most common cancer and the second leading cause of cancer death in men [3]. About 50% of men over 70 years old will experience symptoms such as frequent urination due to BPH, which seriously affects the quality of life. The clinical symptoms and signs of prostate cancer are similar to those of BPH. However, prostate cancer is initially diagnosed at a later clinical stage and has an overall poor prognosis, while the treatment of prostate cancer and benign prostatic hyperplasia is also different. Therefore, in order to improve the therapeutic effect and avoid misdiagnosis or over diagnosis, the differential diagnosis of prostate cancer and prostatic hyperplasia is very important [4]. Clinical pathological and biochemical indicators are important indicators for the diagnosis of prostate cancer

DOI: 10.54254/2755-2721/67/20240697

and benign prostatic hyperplasia. Therefore, this article aims to comprehensively analyze the biochemical and pathological data of prostate tumors, and use a variety of machine learning models to try to establish a good performance prostate classification and prediction model, so as to provide some guidance for clinical treatment.

2. Background

2.1. Current status of clinical diagnosis of prostate cancer

Currently, the "three-step" method is the standard for early clinical prostate cancer diagnosis, involving PSA testing, DRE, imaging (TRUS, MRI), and systematic biopsy. Recommended tests for BPH include history taking, physical examination, IPSS, PSA testing, and ultrasonography.

The Gleason score assesses prostate cancer grade, with higher scores indicating greater likelihood of tumor spread. IPSS categorizes BPH severity based on scores ranging from 0 to 35 [5].

In clinical monitoring and treatment, researchers utilize techniques such as Fuzzy c-means Clustering and CNN algorithms for precise prostate tumor tissue localization on MRI. Previous studies have employed logistic multivariate analysis and machine learning models like RF, SVM, logistic regression, and Naive Bayes to establish prostate cancer diagnosis models. Recent research also explores single-class prediction models (logistic regression, decision tree, SVM) and ensemble methods (random forest, adaptive boosting, extreme gradient boosting) to develop prediction models for prostate cancer and BPH, comparing their performance to identify optimal models [6].

2.2. State of the Art in Machine Learning

Machine learning, an integral part of computer and information science, revolves around the concept of learning, which entails deducing general rules from finite data samples. It encompasses the study of algorithms that enhance their performance based on past experiences [7]. Diverse perspectives exist regarding its implications. Kohavi and Provost, for instance, view machine learning as inductive algorithms applied in the knowledge discovery process, akin to experiential learning described by Shavlik and Dieterich. While Kohavi and Provost emphasize the prior collection of training samples, Shavlik and Dieterich emphasize external provision of samples. In Wilson's machine learning dictionary, machine learning is characterized as a program adjusting itself to yield better outputs for the same inputs in subsequent iterations. Despite challenges in precisely defining machine learning, it is widely acknowledged as a multidisciplinary field incorporating insights from artificial intelligence, probability statistics, computational complexity theory, cybernetics, information theory, philosophy, and neurobiology.

3. Computer experimental data processing and model establishment

This section describes the processing of patient information data as well as the establishment of the model.

3.1. Data Introduction

The patient data were obtained from the Chinese PLA General Hospital Prostate Cancer Warning Dataset V1.0 spanning 2016 to 2021, comprising 200 entries. The dataset is divided into biochemical and pathological components. The biochemical dataset includes age, height, weight, apolipoprotein A Ⅱ, apolipoprotein C2, and other indicators, with labels indicating diagnoses: 1 for prostatic hyperplasia, 2 for prostate cancer, and 3 for both, which we converted to 0. There were 112 BPH cases, 38 prostate cancer cases, and 22 cases of both.

3.2. Data processing

3.2.1. Detect and remove abnormal samples from biochemical data sets. In order to detect outliers in the standardized data, we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DOI: 10.54254/2755-2721/67/20240697

algorithm to detect outliers [8]. The algorithm can replace the clustering algorithms such as KMeans and hierarchical clustering, and divide the area with sufficient density into clusters, find arbitrary shape clusters in the spatial database with noise, and define the cluster as the maximum set of points connected by density.

To detect and remove outliers from unstandardized data, we use an Isolation Forest algorithm to detect outliers: randomly pick a feature value in the dataset, split the dataset into two sets that are greater or less than that feature value, and repeat the process for each of the two sets until the data cannot be separated. In a statistical sense, relatively clustered points need more splits, and relatively isolated points need fewer splits. The isolation forest uses the number of splits to measure whether a point is clustered (normal) or isolated (abnormal).

3.2.2. Identify biochemical dataset outliers and filled with the mean. Outliers were identified by the Inter-Quartile and filled with the mean. The basic idea of the interquartile range algorithm is as follows: First, the positions of the three quartiles that divide the data into four parts are determined, which are set as the lower quartile, the median quartile and the upper quartile, respectively:

$$Q_1=\frac{1*(N+1)}{4} \tag{1}$$

$$Q_2=\frac{2*(N+1)}{4} \tag{2}$$

$$Q_3=\frac{3*(N+1)}{4} \tag{3}$$

$\Delta=Q_3-Q_1$, upper and lower boundary is respectively: $\max(\{x|x\leqslant Q_3+1.5\Delta\})$,$\min(\{x|x\geqslant Q_1-1.5\Delta\})$, as a result, the abnormal value is:$\{x|x>Q_3+1.5\Delta\}$,$\{x|x<Q_1-1.5\Delta\}$.

3.3. Model construction

3.3.1. Combined sampling. Considering the imbalance of data samples, we use SMOTETomek (comprehensive sampling) to combine samples. Oversampling randomly copies a few samples to increase their size, and undersampling randomly undersamples the main classes, so we use SMOTETomek (comprehensive sampling) sampling method.

3.3.2. Feature screening. In order to remove redundant features and irrelevant features and retain useful features, we use RFECV (Recursive Feature Elimination using Cross Validation) and Random Forest for feature selection, and finally take the intersection of the selected features of the two.

figure 1 shows the visualization result of RFECV. It can be seen that when the number of feature values is 22, the score is the highest, so 22 suitable feature variables are selected based on RFECV.

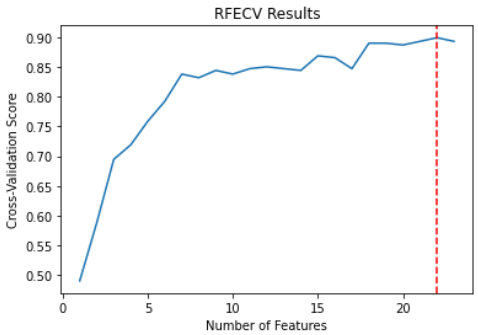

Figure 1. RFECV Visualization result diagram.

figure 2 shows the visualization results of feature importance obtained by random forest. We finally selected features with scores >0.04.

DOI: 10.54254/2755-2721/67/20240697

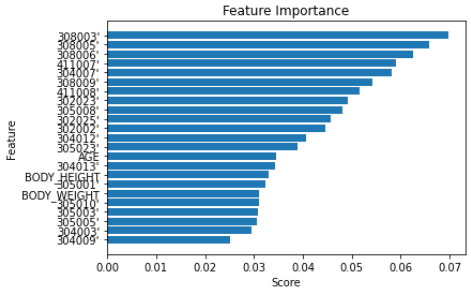

Figure 2. Random forest feature importance map.

The features selected by cross-validation feature elimination are intersected with those selected by random forest, and the features finally selected are: [" 411008 '", "308003", "302002", "411007", "302025", "305008", "308006", "308009", "304012", "' 304007", "308005'", "302023"']. They are PSA (total), lactate dehydrogenase, globulin, PSA (free), serum uric acid serum uric acid, chloride, creatine kinase isoenzyme, alkaline phosphatase, Apolipoprotein A1, high-density lipoprotein cholesterol, creatine kinase and creatinine. A total of 12 characteristics.

3.3.3. Model construction. For the analysis of biochemical data, various algorithms of XGBoost, random forest, LightGBM, CatBoost, and Stacking classifiers are used to attempt the data with filtered eigenvalues, respectively.

For the analysis of pathological data, we adopted two methods.

In the first method, we used CountVectorizer to extract features for new features. Considering the small number of samples, we used ADASYN to over-sample the sampled features, and then used CatBoost, LightGBM and RandomForest to train the sampled features. Finally, the model effect is verified on the test set.

In the second method, the Bert network was trained directly with new features without feature extraction. AdamW was adopted as the optimizer and the loss function was focal loss to improve the model's prediction ability for difficult to classify samples.

4. Experimental results and analysis

4.1. Results and analysis of biochemical data sets

Table 1 shows the results of biochemical data set model classification performance indicators. We can see that for biochemical data set, XGBoost model is not ideal for the classification of only prostate hyperplasia, with an accuracy rate of only 78%. The accuracy of the random forest model for the classification of both prostate hyperplasia and prostate cancer and only prostate cancer can reach 92%, and the accuracy of only prostate hyperplasia and only prostate cancer can reach more than 80%, which is relatively ideal. The LightGBM model was only 72% accurate in classifying only prostate hyperplasia. The CatBoost model was only 70 percent accurate in classifying only prostate cancer. The classification accuracy of the Stacking model can be 96% for both prostate hyperplasia and prostate cancer, 83% for difficult-to-classify BPH, and 88% for prostate cancer. Therefore, the Stacking model is an ideal classification model.

DOI: 10.54254/2755-2721/67/20240697

Table 1. Biochemical data set model classification performance evaluation index

Model	Classification	Precision	Recall	F1-score	Accuracy
XGBoost	both prostate hyperplasia and prostate cancer	0.89	0.96	0.92	0.85
	prostate hyperplasia	0.78	0.78	0.78	
	prostate cancer	0.86	0.78	0.82	
Random forest	both prostate hyperplasia and prostate cancer	0.92	0.92	0.92	0.85
	prostate hyperplasia	0.80	0.67	0.73	
	prostate cancer	0.81	0.91	0.86	
LightGBM	both prostate hyperplasia and prostate cancer	0.92	0.96	0.94	0.83
	prostate hyperplasia	0.72	0.72	0.72	
	prostate cancer	0.82	0.78	0.80	
CatBoost	both prostate hyperplasia and prostate cancer	0.88	0.92	0.90	0.79
	prostate hyperplasia	0.80	0.44	0.57	
	prostate cancer	0.70	0.91	0.79	
Stacking	both prostate hyperplasia and prostate cancer	0.96	0.92	0.94	0.89
	prostate hyperplasia	0.83	0.83	0.83	
	prostate cancer	0.88	0.91	0.89	

4.2. Results and analysis of pathological data sets

The classification accuracy of each model in the pathological data set is shown in table 2. Among the algorithm models based on CountVectorizer and three classifiers, the accuracy of Catboost model and random forest model is less than 80%, and the classification results are not very satisfactory. The LightGBM model performed best with an accuracy of 82%, while the Bert model achieved an accuracy of 80.39%.

Table 2. Accuracy of each model in pathology data set.

Model	Accuracy
CatBoost	78.43%
LightGBM	82.35%
Random forest	74.51%
Model	Accuracy

5. Conclusions

5.1. Results and analysis of biochemical data sets

In order to explore the potential medical significance of establishing the model and to better explain the model, we constructed a feature aggregation diagram to rank the importance of the features, as shown in figure 3 (both prostate hyperplasia and prostate cancer, prostate hyperplasia and prostate cancer in turn). Combined with the analysis of the results, it can be concluded that the corresponding patients should focus on the indicators PSA (total), high-density lipoprotein cholesterol, lactate dehydrogenase, and calcreatine kinase isoenzyme. If indicators of creatine kinase isoenzyme, PSA (total) and HDL cholesterol are abnormal, priority is given to patients with both prostate cancer and hyperplasia. If the indicator globulin, PSA (free) and apolipoprotein A1 are abnormal, it is important to consider that the patient may have prostate hyperplasia. If the indicators PSA (total), apolipoprotein A1, and alkaline phosphatase are abnormal, the patient may be considered to have prostate cancer.

DOI: 10.54254/2755-2721/67/20240697

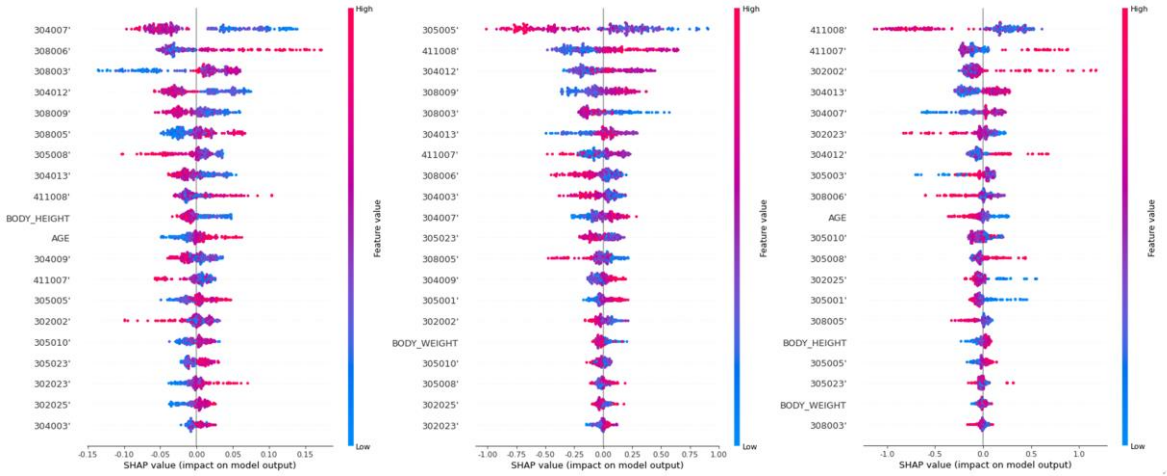

Figure 3. Feature SHAPE map

References

[1] Canovai, M.; Evangelista, M.; Mercatanti, A.; D'Aurizio, R.; Pitto, L.; Marrocolo, F.; Casieri, V.; Pellegrini, M.; Lionetti, V.; Bracarda, S.; Rizzo, M. Secreted miR-210-3p, miR-183-5p and miR-96-5p Reduce Sensitivity to Docetaxel in Prostate Cancer Cells. Cell Death Discov. 2023, 9 (1), 445. https://doi.org/10.1038/s41420-023-01696-4.

[2] S C, S. K.; R, R.; Nachiappa Ganesh, R. Cleistanthins A and B Ameliorate Testosterone-Induced Benign Prostatic Hyperplasia in Castrated Rats by Regulating Apoptosis and Cell Differentiation. Cureus 2022. https://doi.org/10.7759/cureus.32141.

[3] Wang Kunxiang. Prediction Model of Biochemical Recurrence Based on Bayesian Network for Prostate cancer 1 Year after radical surgery [D]. Qingdao University,2022.DOI:10.27262/d.cnki.gqdau.2022.000467

[4] Gui, S.; Lan, M.; Wang, C.; Nie, S.; Fan, B. Application Value of Radiomic Nomogram in the Differential Diagnosis of Prostate Cancer and Hyperplasia. Front. Oncol. 2022, 12, 859625. https://doi.org/10.3389/fonc.2022.859625.

[5] Inoue, L. Y. T.; Trock, B. J.; Partin, A. W.; Carter, H. B.; Etzioni, R. Modeling Grade Progression in an Active Surveillance Study. Stat. Med. 2014, 33 (6), 930–939. https://doi.org/10.1002/sim.6003.

[6] Teng, Y.; Li, W.; Gunasekaran, S. Biosensors Based on Single or Multiple Biomarkers for Diagnosis of Prostate Cancer. Biosens. Bioelectron. X 2023, 15, 100418. https://doi.org/10.1016/j.biosx.2023.100418.

[7] Khan, M. T. I.; Liew, T. W.; Lee, X. Y. Fintech Literacy among Millennials: The Roles of Financial Literacy and Education. Cogent Soc. Sci. 2023, 9 (2), 2281046. https://doi.org/10.1080/23311886.2023.2281046.

[8] Chai Keke. Research on Data cleaning Method of Intelligent Manufacturing of Electric cell [D]. Beijing Jiaotong University,2022.DOI:10.26944/d.cnki.gbfju.2022.001090